The People’s Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet

Christian M. Meyer

Elisabeth Niemann and Iryna Gurevych
International Conference on Computational Semantics
Motivation
Aligning Sense Inventories

Many NLP tasks rely on sense information:

- Word Sense Disambiguation
- Semantic Relatedness
- Machine Translation
- Semantic Search
Motivation
Aligning Sense Inventories

Many NLP tasks rely on sense information:

- Word Sense Disambiguation
- Semantic Relatedness
- Machine Translation
- Semantic Search

WordNet

- precise taxonomy
- textual information
- size
- multilingual
**Motivation**

*Aligning Sense Inventories*

Many NLP tasks rely on sense information:

- Word Sense Disambiguation
- Semantic Relatedness
- Machine Translation
- Semantic Search

---

**WordNet**

- precise taxonomy
- textual information
- size
- multilingual
Motivation
Aligning Sense Inventories

Many NLP tasks rely on sense information:

- Word Sense Disambiguation
- Semantic Relatedness
- Machine Translation
- Semantic Search

Alignment on sense level

- precise taxonomy
- textual information
- size
- multilingual
Alignment on sense level

- **S: (n) damper** (a movable iron plate that regulates the draft in a stove or chimney or furnace)
- **S: (n) damper, muffler** (a device that decreases the amplitude of electronic, mechanical, acoustical, or aerodynamic oscillations)
- **S: (n) damper** (a depressing restraint) “rain put a damper on our picnic plans”

WordNet synset

Wikipedia article
Motivation
Alignment on Sense Level

Alignment on sense level

- **S: (n) damper** (a movable iron plate that regulates the draft in a stove or chimney or furnace)
- **S: (n) damper**, **muffler** (a device that decreases the amplitude of electronic, mechanical, acoustical, or aerodynamic oscillations)
- **S: (n) damper** (a depressing restraint) "rain put a damper on our picnic plans"

WordNet synset  
Wikipedia article
**Motivation**

**Alignment on Sense Level**

- **S:** (n) **damper** (a movable iron plate that regulates the draft in a stove or chimney or furnace)
- **S:** (n) **damper**, **muffler** (a device that decreases the amplitude of electronic, mechanical, acoustical, or aerodynamic oscillations)
- **S:** (n) **damper** (a depressing restraint) "rain put a damper on our picnic plans"

**WordNet synset**

---

**Alignment on sense level**

- **Damper (flow)**
  - From Wikipedia, the free encyclopedia
  - *This article is about the architectural element.*
  - A device that decreases the amplitude of electronic, mechanical, acoustical, or aerodynamic oscillations

- **Damper (food)**
  - From Wikipedia, the free encyclopedia
  - *For other uses of the term "damper", see Damper or similar.*
  - **Damper** is a traditional Australian soda bread prepared in Australia. It is also made in camping situations.
  - Damper was originally developed by stockmen who wanted a bread that could be made without a heated oven.
Motivation
Alignment on Sense Level

Alignment on sense level

Two main benefits:

1. Enhanced sense representation
2. Increase of sense coverage

- S: (n) damper (a movable device that decreases the amplitude of electronic, mechanical, acoustical, or aerodynamic oscillations)
- S: (n) damper, muffler (a device that decreases the amplitude of electronic, mechanical, acoustical, or aerodynamic oscillations)
- S: (n) damper (a depressing restraint) “rain put a damper on our picnic plans”

WordNet synset
Related Work

Automatic Sense Alignment of Wikipedia and WordNet

- Alignment of WordNet and Wikipedia’s **category** system
  - (Suchanek et al., 2007); (Toral et al., 2008/2009); (Ponzetto and Navigli, 2009)
  - Category system is much smaller (0.5M vs. >3M)
  - Neglects huge amount of textual content in articles
  - Different goal: semantically enriched ontology

- Alignment of WordNet and Wikipedia **articles**
  - (Ruiz-Casado et al., 2005): Simple English Wikipedia
  - Alignment based on (normalized) word overlap measure
  - Focus on 1:1 alignment
Related Work

1:1 Alignment vs. n:m Alignment

Both algorithms are modelled in a way that they always align the most likely WordNet synset for a given Wikipedia article (or vice versa):

- What if there is no Wikipedia counterpart for a given WordNet synset (or vice versa)?

  \[ S: (n) \text{dream} \text{ (someone or something wonderful) "this dessert is a dream"} \]

- What if there is more than one Wikipedia article that can be aligned to a WordNet synset (or vice versa)?

  \[ S: (n) \text{photogravure, rotogravure (using photography to produce a plate for printing)} \]
Both algorithms are modelled in a way that they always align the most likely WordNet synset for a given Wikipedia article (or vice versa):

- What if there is no Wikipedia counterpart for a given WordNet synset (or vice versa)?

  \[ S: (n) \text{dream} \ (\text{s}omeone \ or \ \text{something} \ \text{wonderful}) \ “\text{this \ dessert \ is \ a \ dream}” \]

  \[ ?, \quad \text{Need for n:m Alignment!} \]

- What if there is more than one Wikipedia article that can be aligned to a WordNet synset (or vice versa)?

  \[ S: (n) \text{photogravure, rotogravure} \ (\text{using} \ \text{photography \ to \ produce \ a \ plate \ for \ printing}) \]

  \[ \text{Rotogravure} \quad \text{Photogravure} \]

  \[ \text{From \ Wikipedia, \ the \ free \ encyclopedia} \quad \text{From \ Wikipedia, \ the \ free \ encyclopedia} \]

  \[ \text{Rotogravure} \quad \text{Photogravure} \quad \text{is \ an \ intaglio \ printmaking \ or \ photo-mechanical} \]
Aligning Wikipedia and WordNet

Our Contributions

- Novel Two-Step Approach for Sense Alignment
- Well-Balanced Reference Dataset for Evaluation
- Full Alignment Publicly Available
Aligning Wikipedia and WordNet
A Two-Step Approach

1. Candidate extraction
2. Candidate disambiguation
Aligning Wikipedia and WordNet

A Two-Step Approach

1. Candidate extraction
2. Candidate disambiguation
Aligning Wikipedia and WordNet
A Two-Step Approach

1. Candidate extraction
2. Candidate disambiguation

WordNet synset

Wikipedia article 1
Wikipedia article 2
Wikipedia article 3
Wikipedia article 4
Wikipedia article 5
Wikipedia article 6
Wikipedia article ...

Might be one, none, or multiple aligned Wikipedia article
Step 1: Candidate Extraction

Overview

- For each synonymous word in the synset extract
  - Articles with the same title
  - Articles with a matching redirect
  - Articles with an inlink of the form \([\text{target}|\text{label}]\)

Example:

- article *Script (typefaces)*
- article *Script (comics)*
- article *Penmanship* (*Handwriting* has a redirect to *Penmanship*)
- article *Writing System* (*Arabic Alphabet* e.g. links to *Writing System*)

The ‘Arabic alphabet’ is the
\([\text{writing system}|\text{script}]\) used for
writing several languages of ...

(Wolf and Gurevych, 2010)
Step 1: Candidate Extraction

Overview

- For each synonymous word in the synset extract
  - Articles with the same title
  - Articles with a matching redirect
  - Articles with an inlink of the form \[[target|label]\]

Example:

- article *Script (typefaces)*
- article *Script (comics)*
- article *Penmanship* (*Handwriting* has a redirect to *Penmanship*)
- article *Writing System* (*Arabic Alphabet* e.g. links to *Writing System*)

The ‘Arabic alphabet’ is the
[[[writing system|script]]] used for
writing several languages of ...

(Wolf and Gurevych, 2010)

High Recall

→ High Coverage of Alignments
Step 2: Candidate Disambiguation

Overview

- Extract bag-of-words
- Transform them to a vector representation
- Calculate vector similarity scores
- Classify each vector/sense pair as alignment or non-alignment based on a trained threshold
Step 2: Candidate Disambiguation
(a) Bag-of-Words

Synsets are represented by synonyms, gloss, examples
Step 2: Candidate Disambiguation
(b) Vector Representation

Method

bag-of-words

\[
\begin{align*}
0.012 & \\
0.002 & \\
0.085 & \\
& \ldots
\end{align*}
\]

bag-of-words

\[
\begin{align*}
0.002 & \\
0.017 & \\
0.007 & \\
& \ldots
\end{align*}
\]
Step 2: Candidate Disambiguation

(c) Vector Similarity

\[
\text{Sim} = \cos \left( \begin{pmatrix} 0.012 \\ 0.002 \\ 0.085 \\ \ldots \end{pmatrix}, \begin{pmatrix} 0.002 \\ 0.017 \\ 0.007 \\ \ldots \end{pmatrix} \right) = 0.125
\]

or

\[
\text{Sim} = \chi^2 \left( \begin{pmatrix} 0.012 \\ 0.002 \\ 0.085 \\ \ldots \end{pmatrix}, \begin{pmatrix} 0.002 \\ 0.017 \\ 0.007 \\ \ldots \end{pmatrix} \right) = 0.117
\]
Step 2: Candidate Disambiguation

*(d) Alignment Classification*

\[ c(wn, wp) = \begin{cases} 
1 & \text{if } \text{sim}(wn, wp) > t \\
0 & \text{else}, 
\end{cases} \]

- \(t\) is a real valued threshold
- 10-fold cross-validation to determine threshold
- use threshold that maximizes performance
Step 2: Candidate Disambiguation

(b) Vector Representation

![Diagram showing bag-of-words and a method connecting String based (Word Overlap) and Personalized PageRank.]

- Bag-of-words
- Bag-of-words
- Method
- String based (Word Overlap)
  - 0.012
  - 0.002
- Personalized PageRank
  - ...
  - ...
Aligning with Personalized PageRank

**Personalized PageRank**

- PageRank (Brin and Page, 1998) depends on transition probability $c$ and random jump vector $v$
- The initial importance of a vertex can be „personalized“ using random jump vector $v$ (Agirre and Soroa, 2009)
- State of the art in WSD

\[ pr = c \cdot M \cdot pr + (1 - c) \cdot v \]

\[ v_i = \begin{cases} 
\frac{1}{m} & \text{if } i \text{ in bag-of-words} \\
0 & \text{otherwise} 
\end{cases} \]

- Personalization based on our bag-of-words
- Vertices with a word from our bag-of-words receive $1/m$ score
- $m = \text{number of synsets in bag-of-words}$
Aligning with Personalized PageRank

*Our Method: ppr*

vertices = synsets
edges = relations

0.012
0.002
0.085
...

0.002
0.017
0.007
...

WordNet synset

bag-of-words

vertices = synsets
edges = relations

Wikipedia article

bag-of-words

vertices = synsets
edges = relations
Aligning with Personalized PageRank

Our Method: ppr

< plant, flora (a living organism …) >
Aligning with Personalized PageRank

Our Method: $ppr_d$

Variant: initialize the PageRank algorithm solely with the synset

WordNet synset → bag-of-words → vertices = synsets → edges = relations → $ppr$

Wikipedia article → bag-of-words → vertices = synsets → edges = relations → $ppr$
Aligning with Personalized PageRank

Our Method: $ppr_d$

- **WordNet synset**
- **Wikipedia article**
- **Bag-of-words**

Vertices = synsets
Edges = relations

$ppr_d$ values:
- 0.012
- 0.002
- 0.085
- ...

$ppr$ values:
- 0.002
- 0.017
- 0.007
- ...

14.01.2011 | Computer Science Department | UKP Lab – Prof. Dr. Iryna Gurevych | Christian M. Meyer | 29
Gold Standard

Well-Balanced Reference Dataset

- 320 WordNet noun synsets covering:
  - Different synset sizes
  - Different shortest path lengths to root
  - Different unique beginners
  - Different number of extracted Wikipedia article candidates

- 1,815 sense alignment candidates
  - Annotated by three human annotators
  - Good pairwise annotator agreement: \( \kappa = 0.866 \ldots 0.878 \)
  - Gold standard created using majority vote
  - 227 pairs were annotated as alignment
  - 221 synsets could be aligned to at least one Wikipedia article
  - for the remaining 99 synsets, no Wikipedia article could be aligned
Gold Standard
Well-Balanced Reference Dataset

- 320 WordNet noun synsets covering:
  - Different synset sizes
  - Different shortest path lengths to root
  - Different unique beginners
  - Different number of extracted Wikipedia article candidates

- 1,815 sense alignment candidates
  - Annotated by three human annotators
  - Good pairwise annotator agreement: $\kappa = 0.866 \ldots 0.878$
  - Gold standard created using majority vote
  - 227 pairs were annotated as alignment
  - 221 synsets could be aligned to at least one Wikipedia article
  - for the remaining 99 synsets, no Wikipedia article could be aligned
Evaluation
Experimental Setup

- **Baselines** (1:1 alignment)
  
<table>
<thead>
<tr>
<th>Baselines</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>for each synset, select a random Wikipedia candidate</td>
</tr>
<tr>
<td>MFS</td>
<td>for each synset, select the most frequently linked Wikipedia article</td>
</tr>
</tbody>
</table>

- **Bag of words representation – WordNet**
  
<table>
<thead>
<tr>
<th>Representation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYN</td>
<td>synonyms, gloss &amp; example sentence from the synset</td>
</tr>
<tr>
<td>SYN+HYPO</td>
<td>SYN plus representation of all hyponyms</td>
</tr>
<tr>
<td>SYN+HYPER</td>
<td>SYN plus representation of all hypernyms</td>
</tr>
<tr>
<td>SYN+HYP2</td>
<td>SYN plus representation of all hyponyms and hypernyms</td>
</tr>
</tbody>
</table>

- **Bag of words representation – Wikipedia**
  
<table>
<thead>
<tr>
<th>Representations</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>Article title</td>
</tr>
<tr>
<td>P</td>
<td>First paragraph</td>
</tr>
<tr>
<td>R</td>
<td>Redirects</td>
</tr>
<tr>
<td>C</td>
<td>Categories</td>
</tr>
</tbody>
</table>
### Evaluation

**Results (1)**

- Random baseline: 0.527
- MFS baseline: 0.534

All figures refer to $F_1$ measure

<table>
<thead>
<tr>
<th>WordNet</th>
<th>Wikipedia</th>
<th>string</th>
<th>$\text{ppr}_d$</th>
<th>$\text{ppr}_d + \text{string}$</th>
<th>$\text{ppr}$</th>
<th>$\text{ppr} + \text{string}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Evaluation

Results (2)

- Random baseline: 0.527
- MFS baseline: 0.534

<table>
<thead>
<tr>
<th>WordNet</th>
<th>Wikipedia</th>
<th>string</th>
<th>ppr_d</th>
<th>ppr_d + string</th>
<th>ppr</th>
<th>ppr + string</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYN</td>
<td>P+T+C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+HYPO</td>
<td>P+T+C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+HYPER</td>
<td>P+T+C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+HYP2</td>
<td>P+T+C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Inclusion of categories (C) increases performance. Inclusion of redirects (R) decrease performance. P+T+C obtained the best results.

All figures refer to F₁ measure.
### Evaluation

**Results (3)**

- Random baseline: 0.527
- MFS baseline: 0.534

#### Table

<table>
<thead>
<tr>
<th>WordNet</th>
<th>Wikipedia</th>
<th>string</th>
<th>ppr$_d$</th>
<th>ppr$_d$ + string</th>
<th>ppr</th>
<th>ppr + string</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYN</td>
<td>P+T+C</td>
<td>.698</td>
<td>.754</td>
<td></td>
<td>.726</td>
<td></td>
</tr>
<tr>
<td>+HYPO</td>
<td>P+T+C</td>
<td>.702</td>
<td>.739</td>
<td></td>
<td>.722</td>
<td></td>
</tr>
<tr>
<td>+HYPER</td>
<td>P+T+C</td>
<td>.738</td>
<td>.752</td>
<td></td>
<td>.765</td>
<td></td>
</tr>
<tr>
<td>+HYP2</td>
<td>P+T+C</td>
<td>.732</td>
<td>.739</td>
<td></td>
<td>.746</td>
<td></td>
</tr>
</tbody>
</table>

---

**Personalized PageRank always outperforms string overlap approach**

- ppr$_d$ outperforms ppr for SYN and +HYPO
- Hypernym synsets increase performance of ppr
Evaluation
Results (4)

- Random baseline: 0.527
- MFS baseline: 0.534

<table>
<thead>
<tr>
<th>WordNet</th>
<th>Wikipedia</th>
<th>string</th>
<th>ppr(_d)</th>
<th>ppr(_d) + string</th>
<th>ppr</th>
<th>ppr + string</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYN</td>
<td>P+T+C</td>
<td>.698</td>
<td>.754</td>
<td>.756</td>
<td>.726</td>
<td>.743</td>
</tr>
<tr>
<td>+HYPO</td>
<td>P+T+C</td>
<td>.702</td>
<td>.739</td>
<td>.747</td>
<td>.722</td>
<td>.740</td>
</tr>
<tr>
<td>+HYPER</td>
<td>P+T+C</td>
<td>.738</td>
<td>.752</td>
<td>.765</td>
<td>.765</td>
<td>.781</td>
</tr>
<tr>
<td>+HYP2</td>
<td>P+T+C</td>
<td>.732</td>
<td>.739</td>
<td>.757</td>
<td>.746</td>
<td>.769</td>
</tr>
</tbody>
</table>

Combinational approach always yields better performance (due to increasing precision)

all figures refer to F\(_1\) measure
Evaluation
Results (5)

- Random baseline: 0.527
- MFS baseline: 0.534

<table>
<thead>
<tr>
<th></th>
<th>WordNet</th>
<th>Wikipedia</th>
<th>string</th>
<th>ppr&lt;sub&gt;d&lt;/sub&gt;</th>
<th>ppr&lt;sub&gt;d&lt;/sub&gt; + string</th>
<th>ppr</th>
<th>ppr + string</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYN</td>
<td>P+T+C</td>
<td></td>
<td>.698</td>
<td>.754</td>
<td>.756</td>
<td>.726</td>
<td>.743</td>
</tr>
<tr>
<td>+HYPO</td>
<td>P+T+C</td>
<td></td>
<td>.702</td>
<td>.739</td>
<td>.747</td>
<td>.722</td>
<td>.740</td>
</tr>
<tr>
<td>+HYPER</td>
<td>P+T+C</td>
<td></td>
<td>.738</td>
<td>.752</td>
<td>.765</td>
<td>.765</td>
<td>.781</td>
</tr>
<tr>
<td>+HYP2</td>
<td>P+T+C</td>
<td></td>
<td>.732</td>
<td>.739</td>
<td>.757</td>
<td>.746</td>
<td>.769</td>
</tr>
</tbody>
</table>

Combination of ppr and string yields best performance with
WordNet synset + hypernyms
Wikipedia article title + first paragraph + categories

all figures refer to F<sub>1</sub> measure
### Error Analysis

- False positives due to highly related sense alignment candidates, e.g.
  
  *(cottonseed, cottonseed oil) or (insulin shock, insulin shock therapy)*

- False negatives due to very different sense representation, e.g.
  
  *<payment, defrayal, defrayment: the act of paying money>*
  *
  *<Payment: A payment is the transfer of wealth from one party…>*

- **Future work:** Include structural knowledge
Conclusions

Lessons Learned

- Novel two-step approach: **Candidate Extraction** and **Disambiguation**
  - Extraction: high recall
  - Disambiguation: Combination of Personalized PageRank and Word Overlap
  - Evaluation reveals $F_1 = 0.781$ on our well-balanced reference dataset

- With our best setting, we generated a **full alignment**
  - Not a 1:1 alignment as in previous works
  - Resources are partly complementary on sense level
  - Increased amount of knowledge for senses found in both resources

- We believe that the new resource and the enhanced knowledge therein can boost the performance of NLP tasks
  - We already started research on integrating the aligned resource in WSD tasks
Ubiquitous Knowledge Processing

Additional Online Material:
http://www.ukp.tu-darmstadt.de/data/sense-alignment/
Thank you for your attention!

Online Resources and Questions

Additional Online Material:
http://www.ukp.tu-darmstadt.de/data/sense-alignment/
Kontakt / Contact

Christian M. Meyer
Technische Universität Darmstadt
Ubiquitous Knowledge Processing Lab

馥 Hochschulstr. 10, 64289 Darmstadt, Germany
✆ +49 (0)6151 16–7477
/msg +49 (0)6151 16–5455
✉ meyer (at) ukp.informatik.tu-darmstadt.de

Legal Issues
The slides are intended for personal use by the audience of the talk. Photographies, illustrations, tradesmarks, or logos are property of the holder of rights. To avoid any misconceptions, I would strongly recommend to get in touch before reusing or redistributing the slides or any additional material of the talk. The same applies if you consider your rights infringed – please let me know to initiate further clarification.